Statistically Based Model Comparison Techniques

H. T. Banks

Center for Research in Scientific Computation (CRSC)
Center for Quantitative Sciences in Biomedicine (CQSB)
North Carolina State University
Raleigh, NC 27695

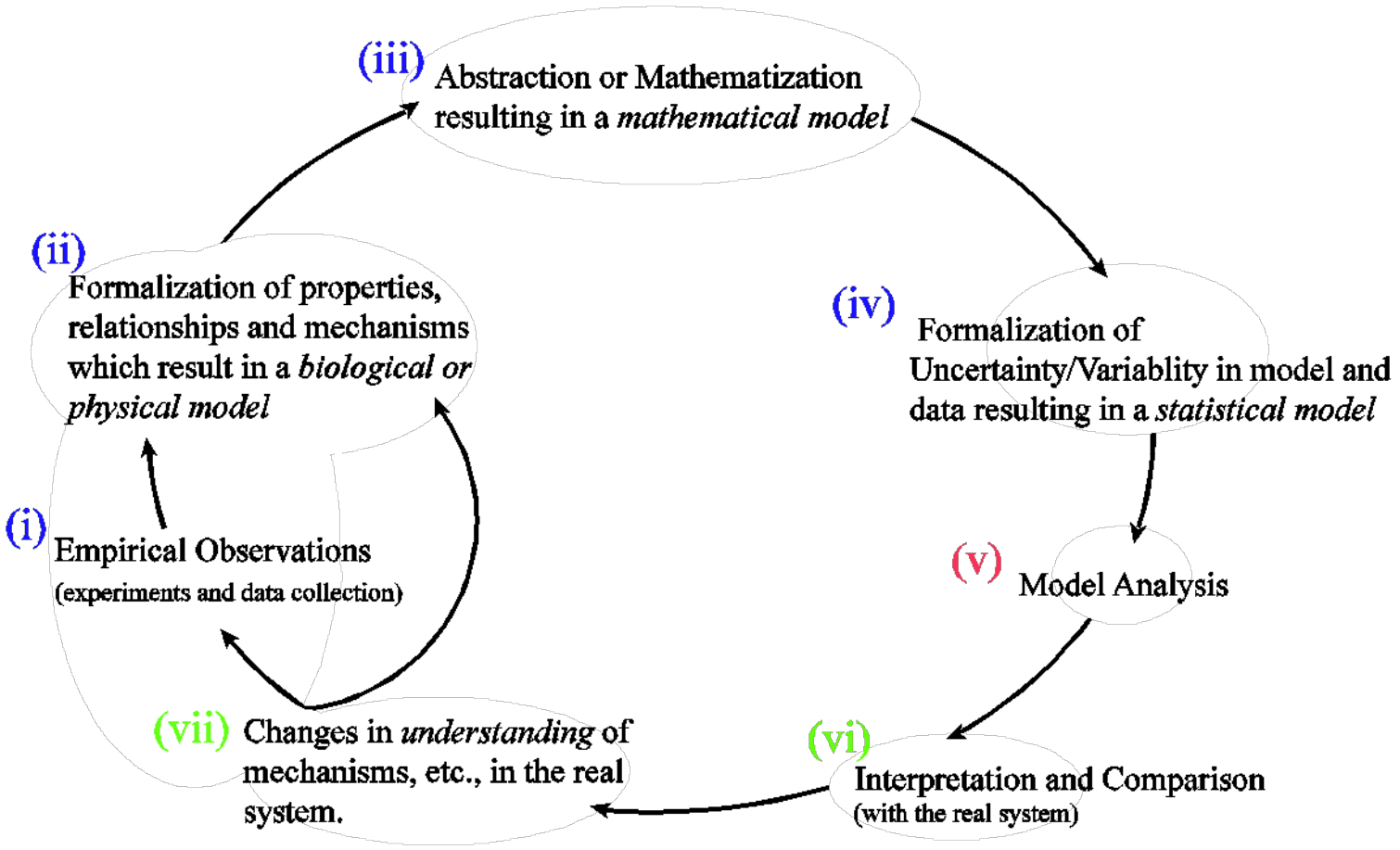
Goals of Modeling

- simplification: use of models for investigation of very complex systems in a systematic manner
- ease in manipulation: separation of subunits and hypothesis testing thru use of simulations in place of experimentation
- assist in formulation of hypotheses and in design of critical experiments
- preciseness: move from general, verbal explanation of phenomena to specific, quantitative one
- organization of inquiry--tends to polarize one's thinking and aid in posing basic questions concerning what one does and does not know for certain about real system
- ***primary goal=enlightenment-***--gain better understanding of real system

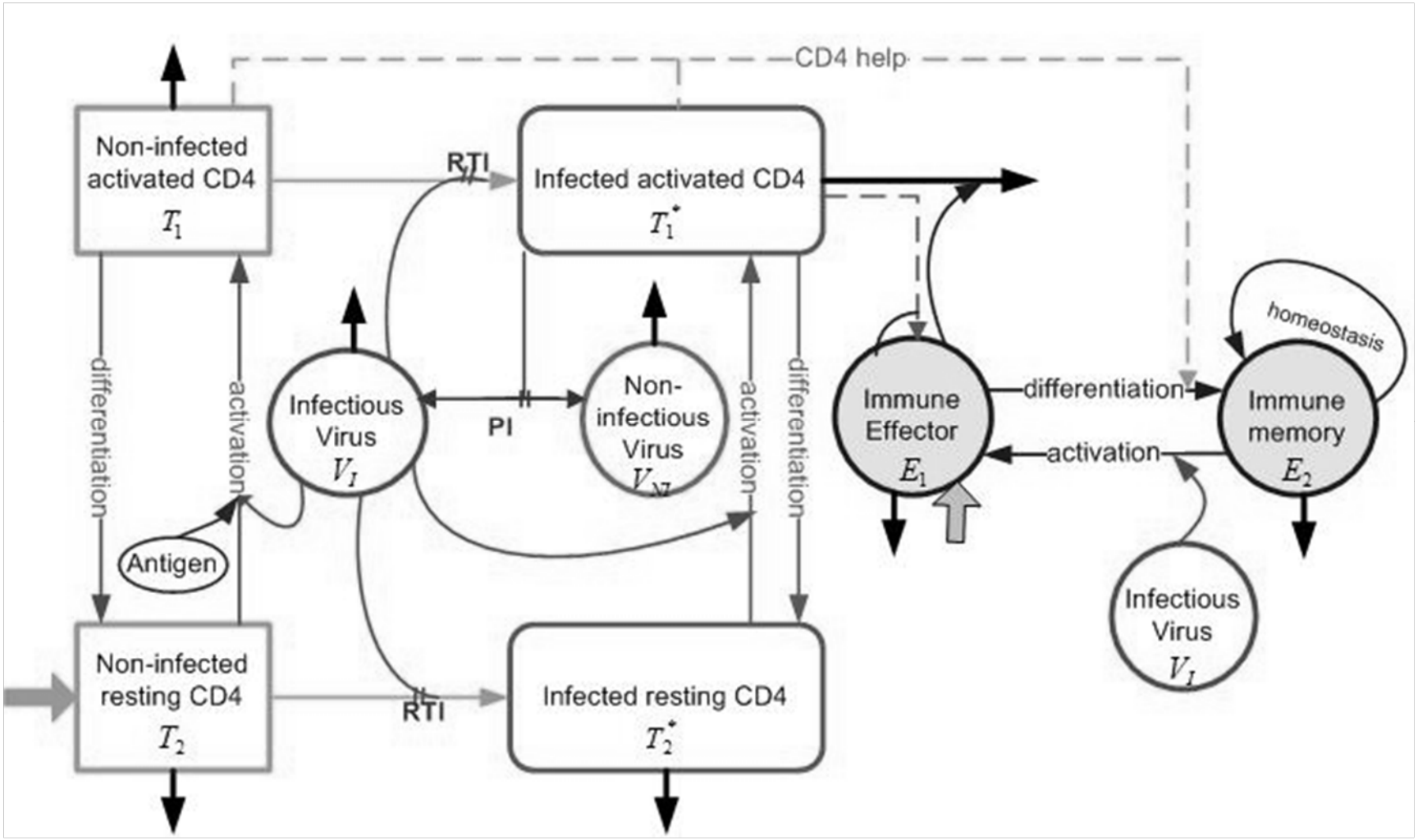
Iterative modeling process

- *Begins* with questions raised by observations or collected data or hypothesized mechanisms
- Mechanism-based relationships in model informed and guided by data
  - Which variables are important in system?
  - Relationships between variables inform nature of terms
- Output of model compared with observations
  - Are they qualitatively/quantitatively similar?
  - Has variance in observations been reasonably modeled? (residual plots)
- Modified model fits to data can be tested for statistical improvement (statistically based model comparison tests)

The Iterative Modeling Process



Formation Stage: (i),(ii),(iii),(iv) Solution Stage: (v) Interpretation Stage: (vi), (vii)


CD4 help
Non-infected activated CD4 T_1
RTI
Infected activated CD4 T_1^*
differentiation
activation
Infectious Virus V_I
PI
Non-infectious Virus V_{NI}
activation
differentiation
Immune Effector E_1
differentiation
activation
homeostasis
Immune memory E_2
Antigen
Infectious Virus V_I
Non-infected resting CD4 T_2
RTI
Infected resting CD4 T_2^*

Statistically Based Model Comparison Techniques

- Previously, discussed techniques (e.g., residual plots) for investigating *correctness* of the assumed *statistical model* underlying the estimation (OLS or GLS) procedures used in inverse problems. To this point have not discussed correctness issues related to choice of *mathematical model*.
- Number of ways in which questions related to mathematical model may arise, e.g, modeling studies [BKa83,BKu89b] can raise questions as to whether a mathematical model can be improved by *more detail* and/or *further refinement*.

- Can we improve mathematical model by assuming more *detail* in a given mechanism (constant rate vs. time or spatially dependent rate) – e.g., see [BBDS]–time dependent mortality rates during sub-lethal damage in insect populations exposed to various levels of pesticides???
- Or one might question whether an *additional mechanism* in model might produce a better fit to data–see [BF1,BF90,BKa83] for *diffusion alone* or *diffusion plus convection* in cat brain transport in grey vs. white matter considerations.
- Does addition of delays yield improved model?? see [BBH]

Before continuing, important point must be made: In model comparison results outlined below, there are really *two models* being compared: the *mathematical model* and the *statistical model*. If one embeds the mathematical model in the *wrong statistical model* (for example, assuming constant variance when this really isn't true), then the mathematical model comparison results using the techniques presented here will be *invalid* (i.e., *worthless*). An important remark in all this is that one must have the mathematical model one wants to simplify or improve (e.g., test whether $\mathcal{V} = 0$ or not in the example below) embedded in the *correct statistical model* (determined in large part by the observation process), so that the comparison actually is *only with regard to the mathematical model*.

Motivation:

- Illustrate with mathematical model for diffusion-convection process–use with experiments to study substance (labelled sucrose) transport in cat brains (heterogeneous–grey and white matter) [BKa83].
- Transport of substance in cat's brains described by PDE (convection/diffusion model) for *change in time and space*:

$$\frac{\partial u}{\partial t} + \mathcal{V}\frac{\partial u}{\partial x} = \mathcal{D}\frac{\partial^2 u}{\partial x^2}. \tag{1}$$

- $\vec{q} = (\mathcal{D}, \mathcal{V}) \in \mathcal{Q} =$ admissible parameter set: $\mathcal{D} =$ diffusion coefficient, $\mathcal{V} =$ bulk velocity of fluid
- Our problem: test whether the parameter $\mathcal{V}$ plays a significant role in the mathematical model.

- If model (1) represents a diffusion-convection process, seek to determine whether diffusion alone or diffusion plus convection best describes transport phenomena represented in cat brain data sets $\{y_{ij}\}$ for $\{u(t_i, x_j; \vec{q})\}$, concentration of labelled sucrose at times $\{t_i\}$ and location $\{x_j\}$.
- Wish to test null hypothesis H_0 that diffusion alone best describes data versus alternative hypothesis H_A that convection also needed–take $H_0 : \mathcal{V} = 0$ and alternative $H_A : \mathcal{V} \neq 0$. Consequently, restricted parameter set $\mathcal{Q}_H \subset \mathcal{Q}$ defined by

$$\mathcal{Q}_H = \{\vec{q} \in \mathcal{Q} : \mathcal{V} = 0\}$$

  important.
- To carry out, need some model comparison tests of *analysis of variance (ANOVA)* type [G76] from statistics involving *residual sum of squares (RSS)* in least squares problems.

RSS Based Statistical Tests

In general, we assume an inverse problem with mathematical model $f(t, \vec{q})$ and n observations $\vec{Y} = \{Y_j\}_{j=1}^n$. We define an OLS performance criterion

$$J_n(\vec{q}) = J_n(\vec{Y}, \vec{q}) = \frac{1}{n} \sum_{j=1}^{n} [Y_j - f(t_j, \vec{q})]^2,$$

where our *statistical model* again has the form

$$Y_j = f(t_j, \vec{q}_0) + \mathcal{E}_j, \quad j = 1, \ldots, n,$$

with $\{\mathcal{E}_j\}_{j=1}^n$ being *independent and identically distributed*, $E(\mathcal{E}_j) = 0$ and *constant variance* $\mathrm{var}(\mathcal{E}_j) = \sigma^2$. As usual $\vec{q}_0$ is the "true" value of $\vec{q}$ which we assume to exist. As noted above, we use $\mathcal{Q}$ to represent the set of all the admissible parameters $\vec{q}$ and assume that $\mathcal{Q}$ is a compact subset of Euclidean space of R^p with $\vec{q}_0 \in \mathcal{Q}$.

Let $q^n(\vec{Y}) = q^n_{OLS}(\vec{Y})$ be the OLS *estimator* using J_n with corresponding *estimate* $\hat{q}^n = q^n_{OLS}(\vec{y})$ for a realization $\vec{y} = \{y_j\}$ so

$$q^n(\vec{Y}) = \arg\min_{\vec{q}\in\mathcal{Q}} J_n(\vec{Y}, \vec{q}) \quad \text{and} \quad \hat{q}^n = \arg\min_{\vec{q}\in\mathcal{Q}} J_n(\vec{y}, \vec{q}).$$

Remark: In most calculations, one actually uses approximation f^N to f (often numerical solution to ODE or PDE for modeling dynamical system)–tacitly assume f^N converges to f–Also questions related to approximations of set $\mathcal{Q}$ when infinite dimensional (e.g., in case of function space parameters such as time or spatially dependent parameters) by finite dimensional discretizations $\mathcal{Q}^M$–see [BKu89b,BF90] for extensive discussions on convergences $f^N \to f$ and $\mathcal{Q}^M \to \mathcal{Q}$–ignore these issues here, keeping in mind these approximations will also be of importance in the methodology discussed below in most practical uses.

In many instances, interested in using data to address whether or not the "true" parameter $\vec{q}_0$ can be found in a subset $\mathcal{Q}_H \subset \mathcal{Q}$, assumed here to be defined by

$$\mathcal{Q}_H = \{\vec{q} \in \mathcal{Q} | H\vec{q} = c\}, \tag{2}$$

H is $r \times p$ matrix of *full rank*, c a known constant vector. Test *null hypothesis* H_0: $\vec{q}_0 \in \mathcal{Q}_H$. Define

$$q_H^n(\vec{Y}) = \arg \min_{\vec{q} \in \mathcal{Q}_H} J_n(\vec{Y}, \vec{q}) \quad \text{and} \quad \hat{q}_H^n = \arg \min_{\vec{q} \in \mathcal{Q}_H} J_n(\vec{y}, \vec{q})$$

and observe that $J_n(\vec{Y}, \hat{q}_H^n) \geq J_n(\vec{Y}, \hat{q}^n)$. Define related non-negative *test statistics* and their *realizations*, respectively, by

$$T_n(\vec{Y}) = n(J_n(\vec{Y}, q_H^n) - J_n(\vec{Y}, q^n)) \text{ and } \hat{T}_n = T_n(\vec{y}).$$

One can establish asymptotic convergence results for the test statistics $T_n(\vec{Y})$–given in detail in [BF90]. These results can, in turn, be used to establish a fundamental result about more useful statistics for model comparison. We define these statistics by

$$U_n(\vec{Y}) = \frac{T_n(\vec{Y})}{J_n(\vec{Y}, q_n)}, \tag{3}$$

with corresponding realizations $\hat{U}_n = U_n(\vec{y})$. We then have asymptotic result that is the basis of ANOVA–type tests.

Under reasonable assumptions (very similar to those required in the asymptotic sampling distribution theory discussed in previous sections (see [BF90,BKu89b, F88, SeWi]) involving regularity and the manner in which samples are taken, one can prove a number of convergence results including:

(i) The estimators q^n converge to $\vec{q}_0$ with probability one as $n \to \infty$;

(ii) If H_0 is true, U_n converges in distribution to $U(r)$ as $n \to \infty$ where $U \sim \chi^2(r)$, a χ^2 distribution with r degrees of freedom, where r is the number of constraints specified by the matrix H.

- Recall that H is the $r \times p$ matrix of full rank defining $\mathcal{Q}_H$ and that random variables *converge in distribution* if their corresponding cumulative distribution functions converge point wise at all points of continuity of the limit cdf.
- An example of the χ^2 density is depicted in Figure 1 where the density for $\chi^2(4)$ (χ^2 with $r = 4$ degrees of freedom) is graphed.

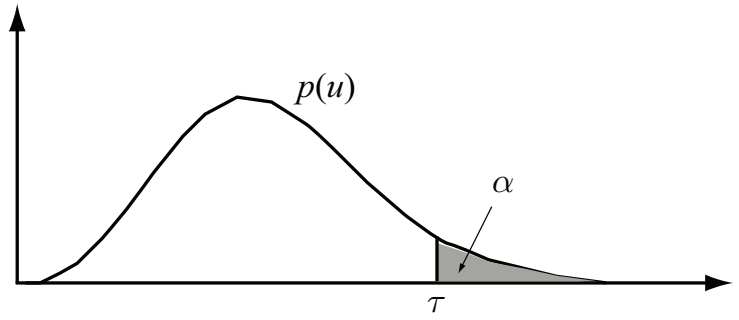


Figure 1: Example of $U \sim \chi^2(4)$ density.

In this figure two parameters (τ, α) of interest are shown. For a given value τ, the value α is simply the probability that the random variable U will take on a value greater than α. That is, $P(U > \tau) = \alpha$ where in hypothesis testing, α is the *significance level* and τ is the *threshold*.

We wish to use this distribution to test the null hypothesis, H_0, which we approximate by $U_n \sim \chi^2(r)$. If the test statistic, $\hat{U}_n > \tau$, then we *reject* H_0 as false with *confidence level* $(1-\alpha)100\%$. Otherwise, we *do not reject* H_0 as true. We emphasize that care should be taken in stating conclusions: we either reject or do not reject H_0 at the specified level of confidence. For the cat brain problem, we use a $\chi^2(1)$ table, which can be found in any elementary statistics text or online and is given here for illustrative purposes, see Table 1.

Table 1: $\chi^2(1)$ values.

α	τ	confidence
.25	1.32	75%
.1	2.71	90%
.05	3.84	95%
.01	6.63	99%
.001	10.83	99.9%

P-Values

The minimum value α^* of α at which H_0 can be rejected is called the *p-value*. Thus, the smaller the p-value, the stronger the evidence in the data in support of rejecting the null hypothesis and including the term in the model, i.e., the more likely the term should be in the model. We implement this as follows: Once we compute $\hat{U}_n = \bar{\tau}$, then $p = \alpha^*$ is the value that corresponds to $\bar{\tau}$ on a χ^2 graph and so we reject the null hypothesis at any confidence level c, such that $c < 1 - \alpha^*$. For example, if for a computed $\bar{\tau}$ we find $p = \alpha^* = .0182$, then we would *reject* H_0 at confidence level $(1 - \alpha^*)100\% = 98.18\%$ or lower. For more information, the reader can consult ANOVA discussions in any good statistics book.

Alternative statement

To test the null hypothesis H_0, we choose a significance level α and use χ^2 tables to obtain the corresponding threshold $\tau = \tau(\alpha)$ so that $P(\chi^2(r) > \tau) = \alpha$. We next compute $\hat{U}_n = \overline{\tau}$ and compare it to τ. If $\hat{U}_n > \tau$, then we *reject* H_0 as false; otherwise, we *do not reject the null hypothesis* H_0.

Application: Cat-Brain Diffusion/Convection Problem

We summarize use of the model comparison techniques outlined above by returning to the cat brain example discussed in detail in [BKa83,BKu89b]. There were *3 sets of experimental data* examined, under the null-hypothesis $H_0 : \mathcal{V} = 0$. For *Data Set 1*, we found after carrying out the inverse problems over $\mathcal{Q}$ and $\mathcal{Q}_H$, respectively,

$$J_n(\hat{q}^n) = 106.15 \quad \text{and} \quad J_n(\hat{q}^n_H) = 180.1.$$

In this case $\hat{U}_n = 5.579$ (note that $n = 8 \neq \infty$), for which $p = \alpha^* = .0182$. Thus, we *reject* H_0 in this case at *any* confidence level less than 98.18%. Thus, we should *reject* that $\mathcal{V} = 0$, which suggests convection is important in describing this data set.

For *Data Set 2*, we found

$$J_n(\hat{q}^n) = 14.68 \quad \text{and} \quad J_n(\hat{q}^n_H) = 15.35,$$

and thus, in this case, we have $\hat{U}_n = .365$, which implies we *do not reject* H_0 with *high degrees of confidence* (p-value very high). This suggests $\mathcal{V} = 0$, which is completely opposite to the findings for Data Set 1.

For the final set (*Data Set 3*) we found

$$J_n(\hat{q}^n) = 7.8 \quad \text{and} \quad J_n(\hat{q}^n_H) = 146.71,$$

which yields in this case, $\hat{U}_n = 15.28$. This, as in the case of the first data set, suggests (with $p < .001$) that $\mathcal{V} \neq 0$ is important in modeling the data.

The difference in conclusions between the first and last sets and that of the second set is interesting and perhaps at first puzzling. However, when discussed with the doctors who provided the data, it was discovered that the first and last set were taken from the *white matter* of the brain, while the other was taken from the *grey matter*. This later finding was consistent with observed microscopic tests on the various matter (micro channels in white matter that promote convective "flow"). Thus, it can be suggested with a reasonably high degree of confidence, that white matter exhibits convective transport, while grey matter does not.

References

[BBDS] H. T. Banks, J.E. Banks, L.K. Dick and J.D. Stark, Estimation of dynamic rate parameters in insect populations undergoing sublethal exposure to pesticides, CRSC-TR05-22, May, 2005; *Bulletin of Mathematical Biology*, **69**, 2007, pp. 2139-2180.

[BBH] H. T. Banks, D.M. Bortz and S.E. Holte, Incorporation of variability into the modeling of viral delays in HIV infection dynamics, CRSC-TR01-25, September, 2001; Revised, November, 2001; *Math Biosci.*, **183** (2003), pp. 63-91.

References

[BDSS] H.T. Banks, M. Davidian, J.R. Samuels, Jr., and K.L. Sutton, An Inverse Problem Statistical Methodology Summary, *CRSC-TR08-01*, January, 2008; Chapter 11 in *Statistical Estimation Approaches in Epidemiology*, (edited by Gerardo Chowell, et al.), Springer, Berlin Heidelberg New York, 2009, pp. 249–302.

[BDE07b] H. T. Banks, S. Dediu and S.E. Ernstberger, Sensitivity functions and their uses in inverse problems, *J. Inverse and Ill-posed Problems*, **15**, 2007, pp. 683-708.

[BEG] H. T. Banks, S.L. Ernstberger and S.L. Grove, Standard errors and confidence intervals in inverse problems: Sensitivity and associated pitfalls,*J. Inv. Ill-posed Problems*, **15**, 2006, pp. 1-18.

[BF1] H. T. Banks and B. G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CCS Rep. 88-16, July, 1988, Brown University; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Note in Biomath., **81**, 1989, pp. 262-273.

[BF90] H. T. Banks and B. G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, September, 1989, University of Southern California; *J. Math. Biol.*, **28**, 1990, pp. 501-527.

References

[BKa83] H. T. Banks and P. Kareiva, Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models, *J. Math. Biol.*, **17**, 1983, pp. 253-272.

[BKu89b] H. T. Banks and K. Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhäuser, Boston, 1989.

[CR] R. J. Carroll and D. Ruppert, *Transformation and Weighting in Regression*, Chapman & Hall, New York, 1988.

References

[CB] G. Casella and R. L. Berger, *Statistical Inference,* Duxbury, California, 2002.

[DG] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.

[F88] B. G. Fitzpatrick, *Statistical Methods in Parameter Identification and Model Selection*, Ph.D. Thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1988.

References

[G] A. R. Gallant, *Nonlinear Statistical Models*, Wiley, New York, 1987.

[G76] F. Graybill, *Theory and Application of the Linear Model*, Duxbury, North Scituate, MA, 1976.

[J] R. I. Jennrich, Asymptotic properties of non-linear least squares estimators, *Ann. Math. Statist.*, **40**, 1969, pp. 633–643.

[Kot] M. Kot, *Elements of Mathematical Ecology*, Cambridge University Press, Cambridge, 2001.

[SeWi] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, J. Wiley & Sons, Hoboken, NJ, 2003.